Analyse en composantes principales (ACP)

François Husson

Laboratoire de mathématiques appliquées - Agrocampus Rennes

husson@agrocampus-ouest.fr

Analyse en Composantes Principales (ACP)

1 Données - Exemples

2 Etude des individus

3 Etude des variables

4 Aides à l'interprétation

Quel type de données ?

L'ACP s'intéresse à des tableaux de données rectangulaires avec des individus en lignes et des variables quantitatives en colonnes

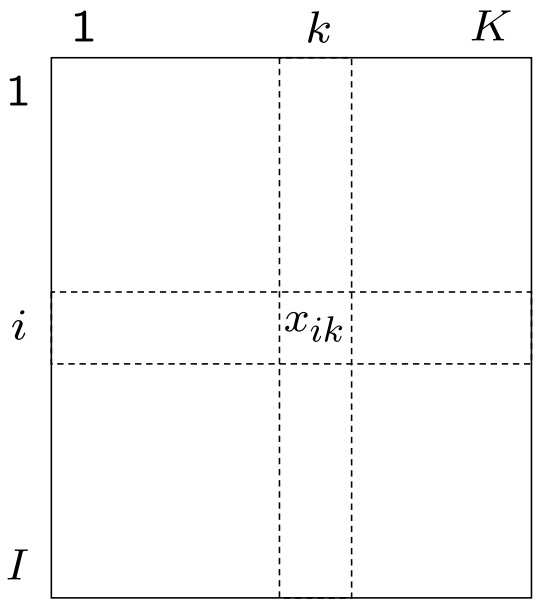


FIGURE: Tableau de données en ACP

Pour la variable k, on note :

$$\text{la moyenne : } \bar{x}_k = \frac{1}{I}\sum_{i=1}^{I} x_{ik}$$

$$\text{l'écart-type : } s_k = \sqrt{\frac{1}{I}\sum_{i=1}^{I}(x_{ik} - \bar{x}_k)^2}$$

Exemples

- Analyse sensorielle : note du descripteur k pour le produit i
- Ecologie : concentration du polluant k dans la rivière i
- Economie : valeur de l'indicateur k pour l'année i
- Génétique : expression du gène k pour le patient i
- Biologie : mesure k pour l'animal i
- Marketing : valeur d'indice de satisfaction k pour la marque i
- Sociologie : temps passé à l'activité k par les individus de la CSP i
- etc.

$\Rightarrow$ Il existe de très nombreux tableaux comme cela

Les données température

- 15 individus (lignes) : villes de France
- 14 variables (colonnes) :
  - 12 températures mensuelles moyennes (sur 30 ans)
  - 2 variables géographiques (latitude, longitude)

	Janv	Févr	Mars	Avri	Mai	Juin	juil	Août	Sept	Octo	Nove	Déce	Lati	Long
Bordeaux	5.6	6.6	10.3	12.8	15.8	19.3	20.9	21	18.6	13.8	9.1	6.2	44.5	-0.34
Brest	6.1	5.8	7.8	9.2	11.6	14.4	15.6	16	14.7	12	9	7	48.24	-4.29
Clermont	2.6	3.7	7.5	10.3	13.8	17.3	19.4	19.1	16.2	11.2	6.6	3.6	45.47	3.05
Grenoble	1.5	3.2	7.7	10.6	14.5	17.8	20.1	19.5	16.7	11.4	6.5	2.3	45.1	5.43
Lille	2.4	2.9	6	8.9	12.4	15.3	17.1	17.1	14.7	10.4	6.1	3.5	50.38	3.04
Lyon	2.1	3.3	7.7	10.9	14.9	18.5	20.7	20.1	16.9	11.4	6.7	3.1	45.45	4.51
Marseille	5.5	6.6	10	13	16.8	20.8	23.3	22.8	19.9	15	10.2	6.9	43.18	5.24
Montpellier	5.6	6.7	9.9	12.8	16.2	20.1	22.7	22.3	19.3	14.6	10	6.5	43.36	3.53
Nantes	5	5.3	8.4	10.8	13.9	17.2	18.8	18.6	16.4	12.2	8.2	5.5	47.13	-1.33
Nice	7.5	8.5	10.8	13.3	16.7	20.1	22.7	22.5	20.3	16	11.5	8.2	43.42	7.15
Paris	3.4	4.1	7.6	10.7	14.3	17.5	19.1	18.7	16	11.4	7.1	4.3	48.52	2.2
Rennes	4.8	5.3	7.9	10.1	13.1	16.2	17.9	17.8	15.7	11.6	7.8	5.4	48.05	-1.41
Strasbourg	0.4	1.5	5.6	9.8	14	17.2	19	18.3	15.1	9.5	4.9	1.3	48.35	7.45
Toulouse	4.7	5.6	9.2	11.6	14.9	18.7	20.9	20.9	18.3	13.3	8.6	5.5	43.36	1.26
Vichy	2.4	3.4	7.1	9.9	13.6	17.1	19.3	18.8	16	11	6.6	3.4	46.08	3.26

Problèmes - objectifs

Le tableau peut être vu comme un ensemble de lignes ou un ensemble de colonnes

Etude des individus

- Quand dit-on que 2 individus se ressemblent du point de vue de l'ensemble des variables ?
- Si beaucoup d'individus, peut-on faire un bilan des ressemblances ?

$\Rightarrow$ construction de groupes d'individus, partition des individus

Problèmes - objectifs

Etude des variables

- Recherche des ressemblances entre variables
- Entre variables, on parle plutôt de liaisons
- Liaisons linéaires sont simples, très fréquentes et résument de nombreuses liaisons $\Rightarrow$ coefficient de corrélation

$\Rightarrow$ visualisation de la matrice des corrélations
$\Rightarrow$ recherche d'un petit nombre d'indicateurs synthétiques pour résumer beaucoup de variables (ex. d'indicateur synthétique a priori : la moyenne, mais ici on recherche des indicateurs synthétiques a posteriori, à partir des données)

Problèmes - objectifs

Lien entre les deux études

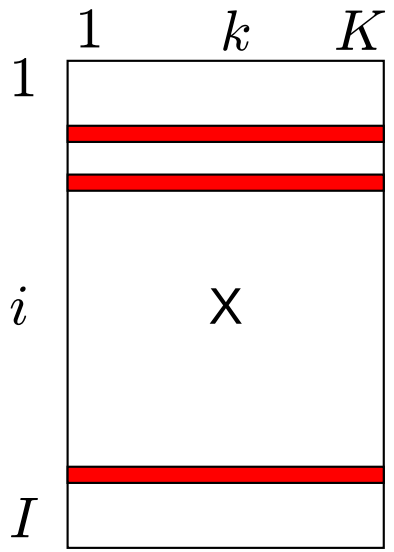
- Caractérisation des classes d'individus par les variables
  $\Rightarrow$ besoin de procédure automatique
- Individus spécifiques pour comprendre les liaisons entre variables
  $\Rightarrow$ utilisation d'individus extrêmes (en terme de variables : langage abstrait mais puissant, revenir aux individus pour voir les choses plus simplement)

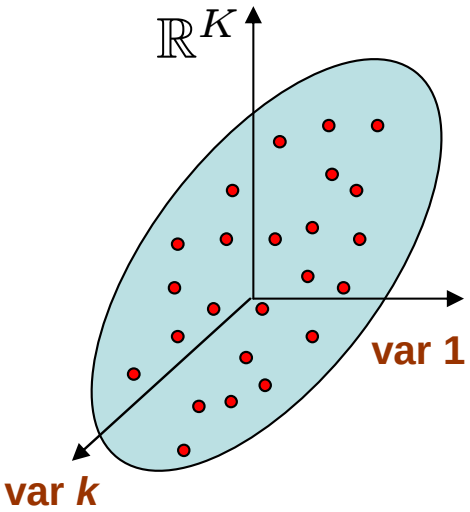
Objectifs de l'ACP :

- Descriptif - exploratoire : visualisation de données par graphiques simples
- Synthèse - résumé de grands tableaux individus $\times$ variables

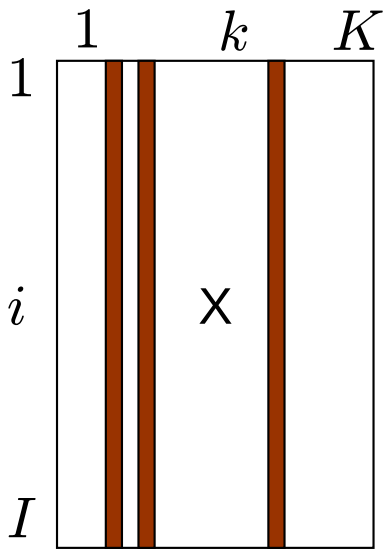
Deux nuages de points

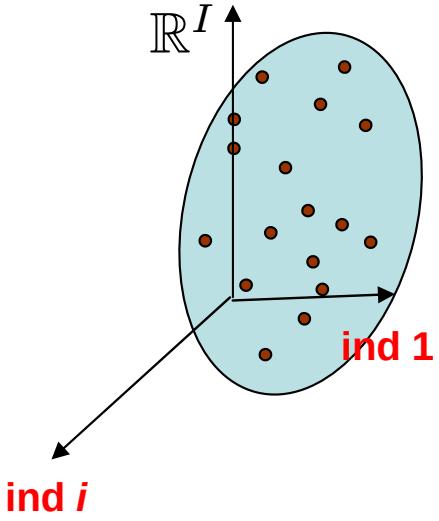
Etude des individus





Etude des variables





Analyse en Composantes Principales (ACP)

1 Données - Exemples

2 Etude des individus

3 Etude des variables

4 Aides à l'interprétation

Le nuage des individus N^I

1 individu = 1 ligne du tableau $\Rightarrow$ 1 point dans un espace à K dim

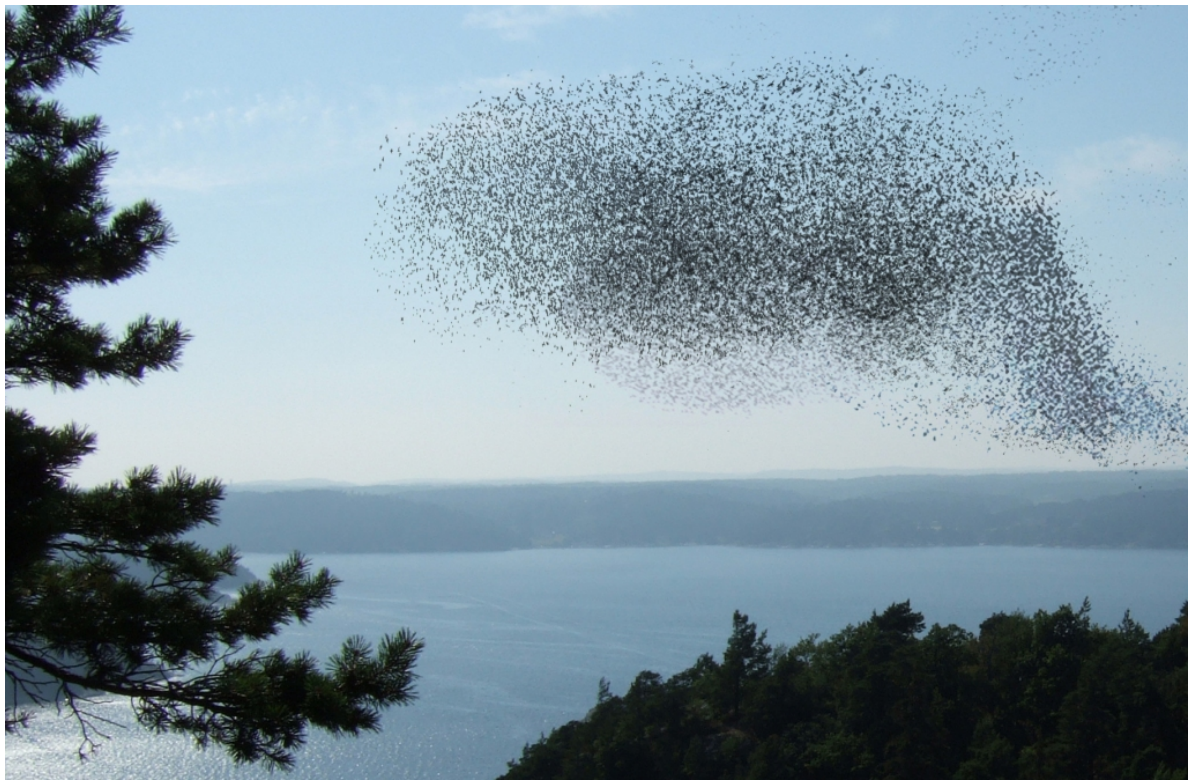
- Si $K = 1$: Représentation axiale
- Si $K = 2$: Nuage de points
- Si $K = 3$: Représentation + difficile en 3D
- Si $K = 4$: Impossible à représenter MAIS le concept est simple

Notion de ressemblance : distance (au carré) entre individus i et i' :

$$d^2(i, i') = \sum_{k=1}^{K} (x_{ik} - x_{i'k})^2 \qquad \text{(merci Pythagore)}$$

Etude des individus $\equiv$ Etude de la forme du nuage N^I

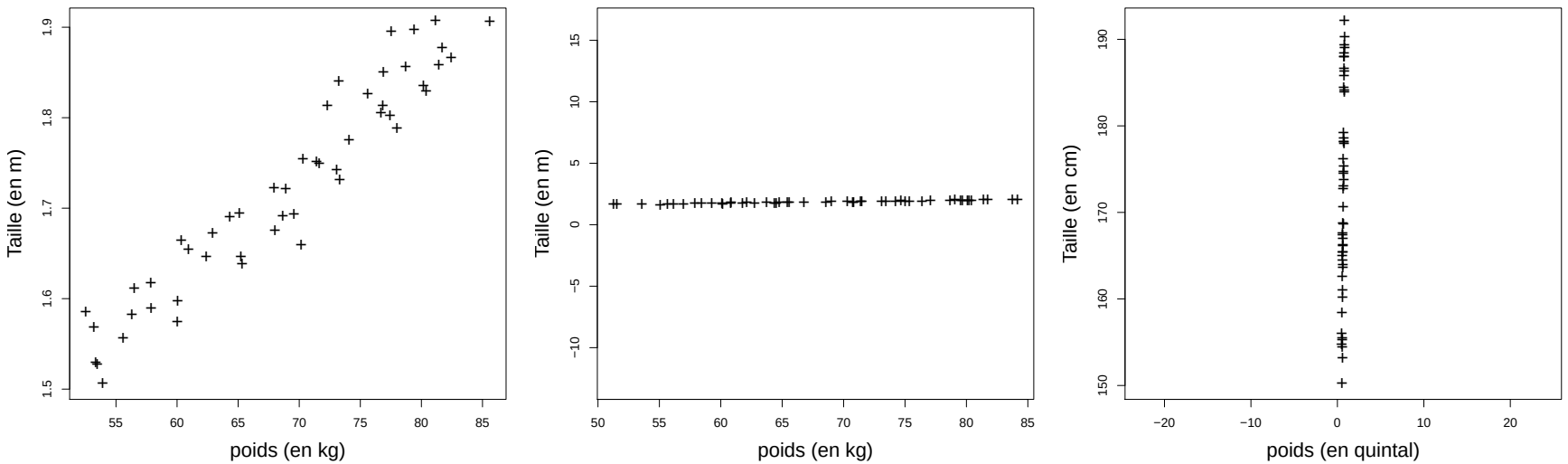
Le nuage des individus N^I



- Etudier la structure, *i.e.* la forme du nuage des individus
- Les individus vivent dans $\mathbb{R}^K$

Centrage – réduction des données

- Centrer les données ne modifie pas la forme du nuage
  $\Rightarrow$ toujours centrer



- Réduire les données est indispensable si les unités de mesure sont différentes d'une variable à l'autre

$$x_{ik} \hookrightarrow \frac{x_{ik} - \bar{x}_k}{s_k}$$

Centrage – réduction des données

	Janv	Févr	Mars	Avri	Mai	Juin	juil	Août	Sept	Octo	Nove	Déce
Bordeaux	0.84	0.98	1.40	1.33	0.94	0.85	0.52	0.74	0.90	0.84	0.67	0.72
Brest	1.10	0.54	-0.29	-1.30	-1.95	-1.98	-2.06	-1.83	-1.28	-0.18	0.62	1.14
Clermont	-0.71	-0.63	-0.50	-0.50	-0.44	-0.31	-0.21	-0.24	-0.44	-0.63	-0.76	-0.66
Grenoble	-1.28	-0.90	-0.36	-0.28	0.05	-0.02	0.13	-0.03	-0.16	-0.52	-0.82	-1.35
Lille	-0.81	-1.07	-1.51	-1.52	-1.40	-1.46	-1.33	-1.27	-1.28	-1.09	-1.05	-0.71
Lyon	-0.97	-0.85	-0.36	-0.06	0.32	0.38	0.42	0.27	-0.05	-0.52	-0.70	-0.92
Marseille	0.79	0.98	1.20	1.48	1.63	1.71	1.69	1.66	1.63	1.52	1.30	1.09
Montpellier	0.84	1.03	1.13	1.33	1.22	1.31	1.39	1.41	1.30	1.29	1.19	0.87
Nantes	0.53	0.26	0.11	-0.13	-0.37	-0.37	-0.50	-0.50	-0.33	-0.07	0.16	0.35
Nice	1.82	2.03	1.74	1.70	1.56	1.31	1.39	1.51	1.86	2.08	2.05	1.77
Paris	-0.30	-0.41	-0.43	-0.20	-0.09	-0.19	-0.36	-0.45	-0.55	-0.52	-0.47	-0.29
Rennes	0.43	0.26	-0.23	-0.64	-0.92	-0.94	-0.94	-0.91	-0.72	-0.41	-0.07	0.29
Strasbourg	-1.84	-1.85	-1.78	-0.86	-0.30	-0.37	-0.41	-0.65	-1.06	-1.60	-1.74	-1.87
Toulouse	0.37	0.42	0.65	0.45	0.32	0.50	0.52	0.69	0.74	0.55	0.39	0.35
Vichy	-0.81	-0.79	-0.77	-0.79	-0.57	-0.42	-0.26	-0.39	-0.55	-0.75	-0.76	-0.76

ACP $\equiv$ Analyse du tableau centré-réduit

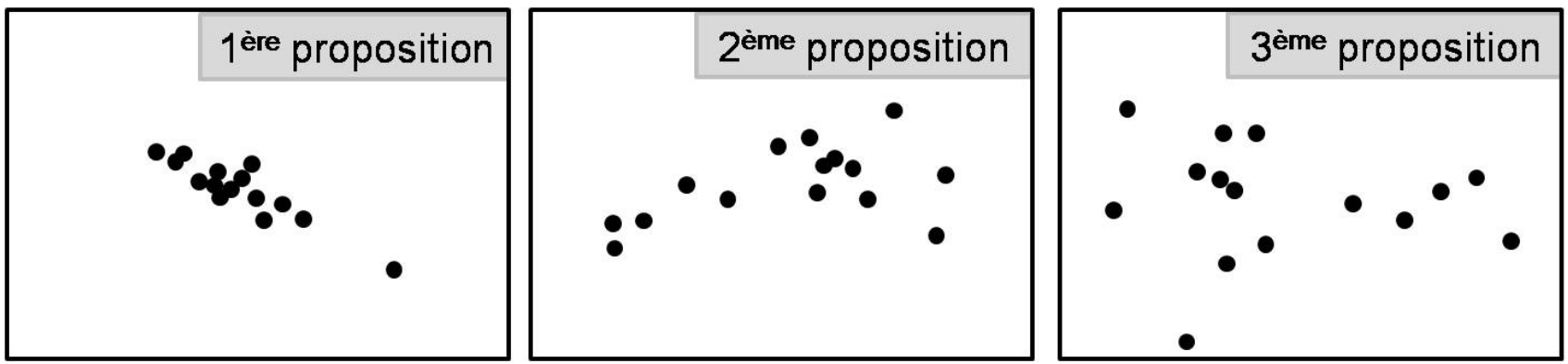
Difficile de voir le nuage $N^I \Rightarrow$ on essaie d'en avoir une image approchée

Ajustement du nuage des individus

L'ACP vise à fournir une image simplifiée de N^I la + fidèle possible
$\Longleftrightarrow$ Trouver le sous-espace qui résume au mieux les données

Qualité d'une image :

- Restitue fidèlement la forme générale du nuage (*animation*)



Ajustement du nuage des individus

L'ACP vise à fournir une image simplifiée de N^I la + fidèle possible
$\Longleftrightarrow$ Trouver le sous-espace qui résume au mieux les données

Qualité d'une image :

- Restitue fidèlement la forme générale du nuage (*animation*)
- Meilleure représentation de la diversité, de la variabilité
- Ne perturbe pas les distances entre individus

Comment quantifier la qualité d'une image ?

A l'aide de la notion de dispersion ou variabilité appelée **Inertie**

Inertie $\equiv$ variance généralisée à plusieurs dimensions

Ajustement du nuage des individus

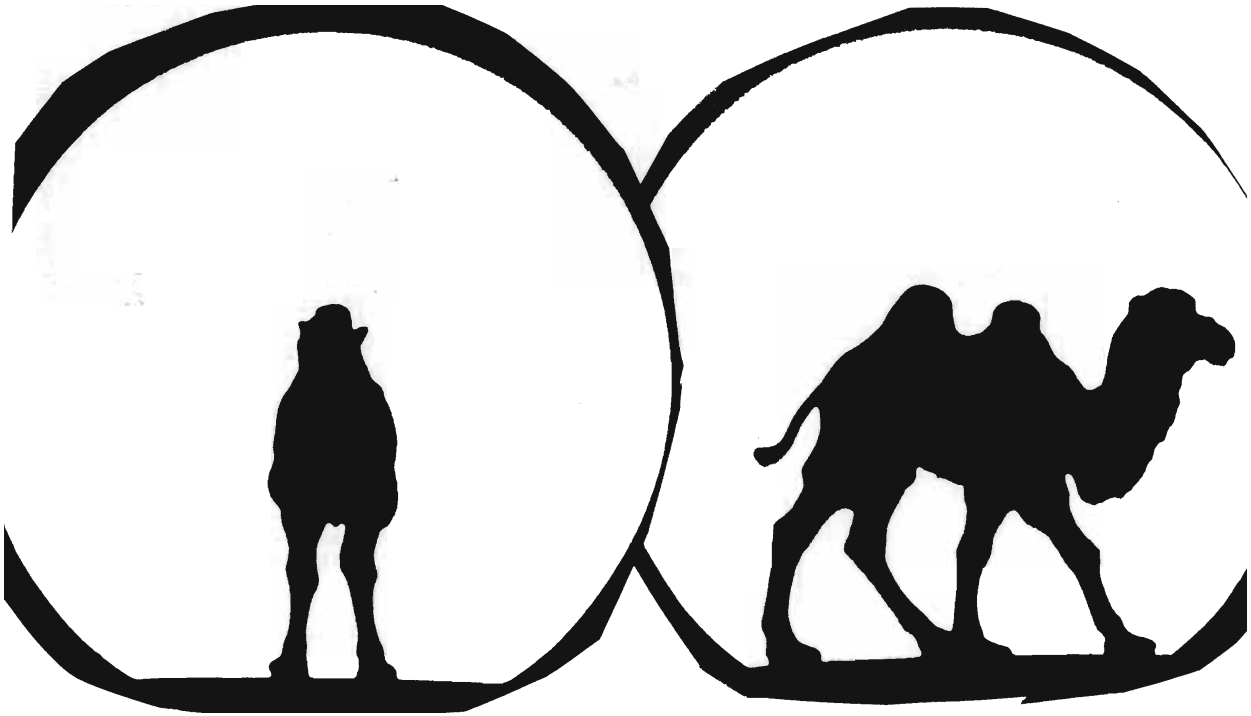
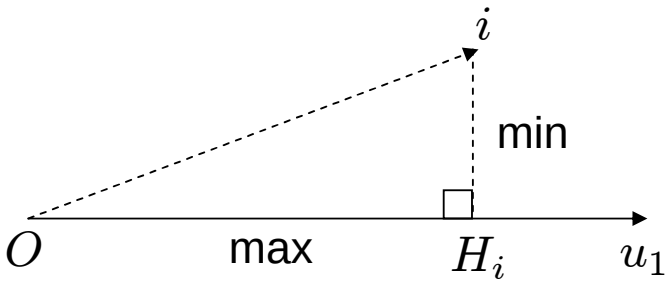


FIGURE: Quel animal ? (*illustration JP Fénelon*)

Ajustement du nuage des individus

Comment trouver la meilleure image approchée du nuage ?

1. Trouver l'axe (facteur) qui déforme le moins possible le nuage

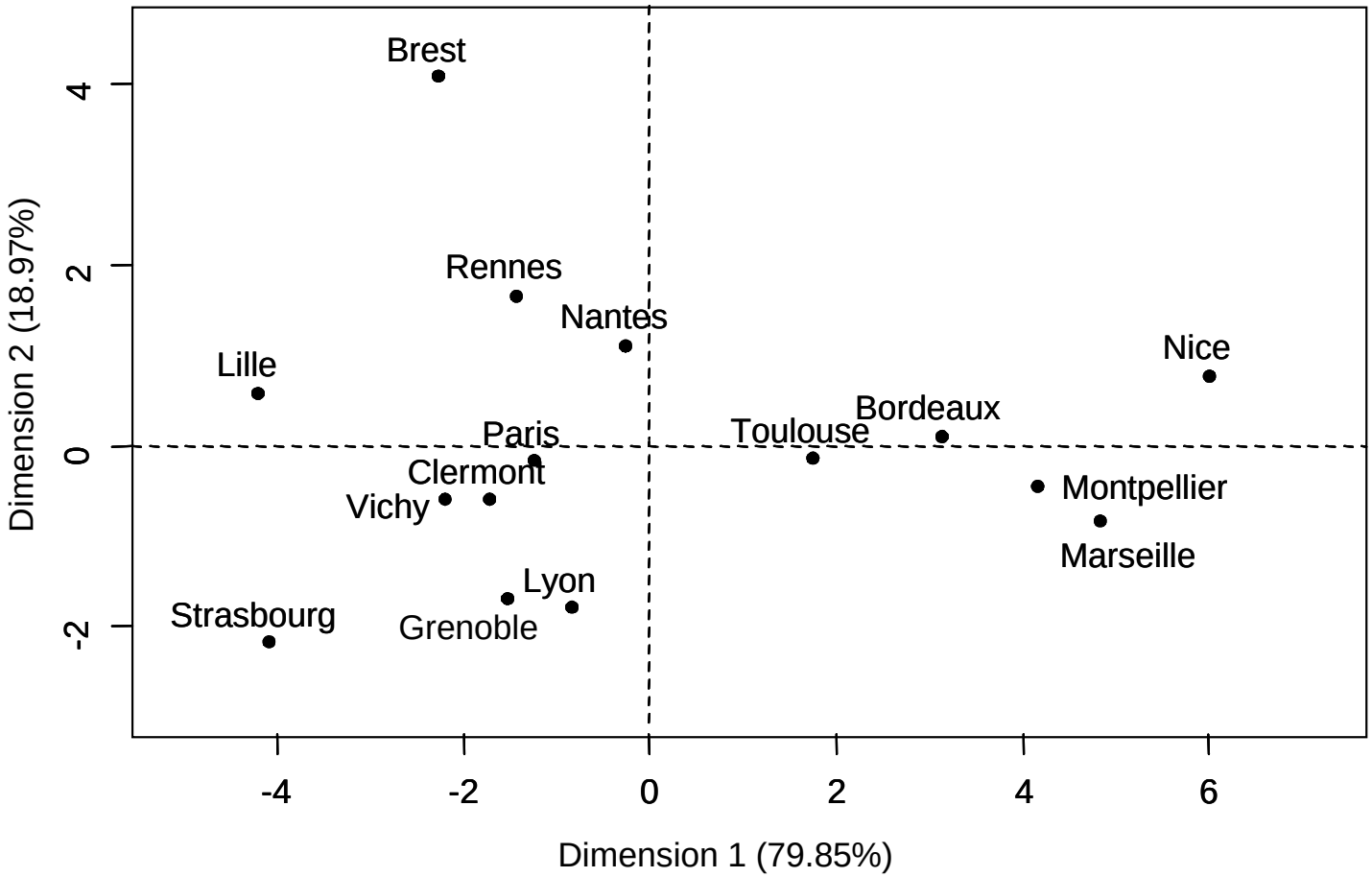


$(iH_i)^2$ petit avec $H_i \in$ axe $\Leftrightarrow$
$(OH_i)^2$ grand (Pythagore)
$\Rightarrow$ on veut $\sum_i (OH_i)^2$ grand

2. Trouver le meilleur plan : maximiser $\sum_i (OH_i)^2$ avec $H_i \in$ plan
Meilleur plan contient le meilleur axe : on cherche $u_2 \perp u_1$ et maximisant $\sum_i (OH_i)^2$

3. on peut chercher un 3ème axe, etc. d'inertie maximum

Exemple : graphe des individus

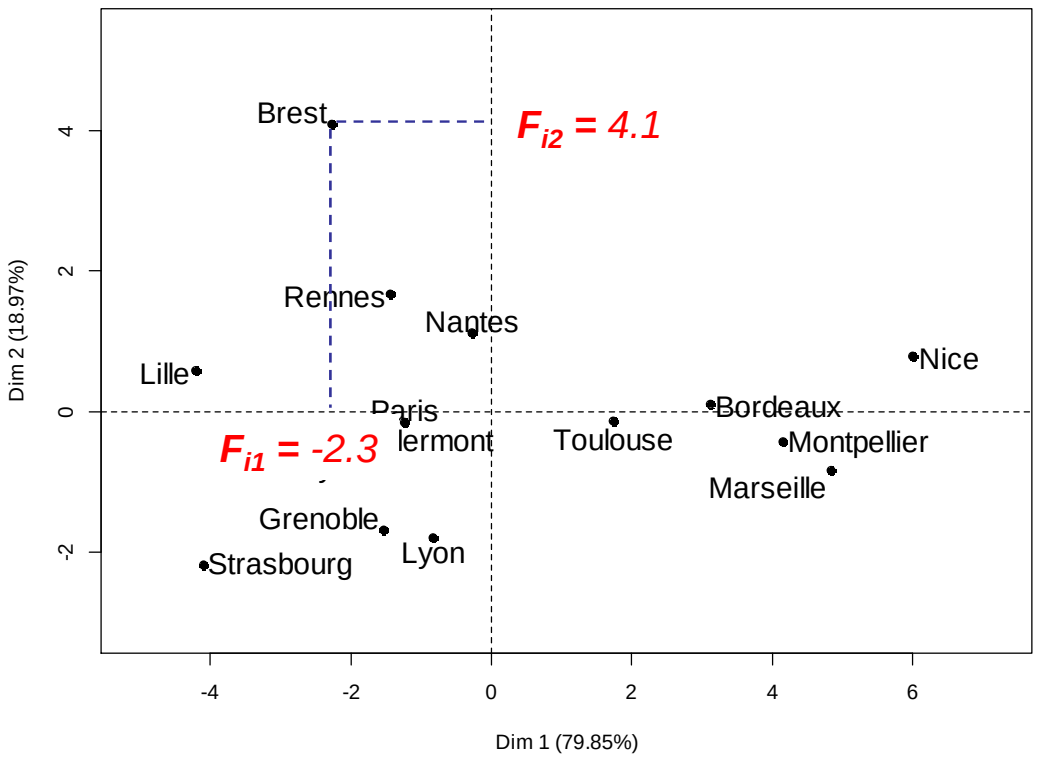


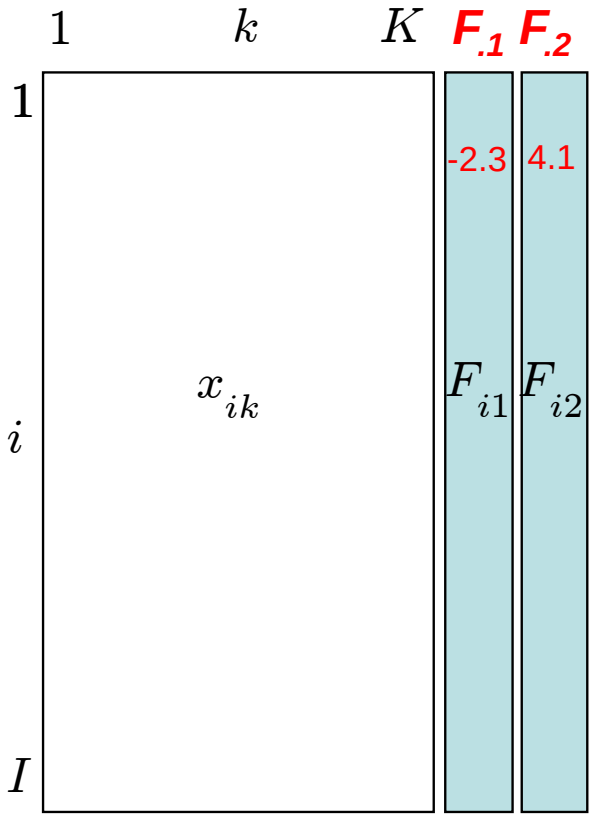
Comment interpréter les axes ? Qu'est-ce qui oppose Lille à Nice ?

$\Rightarrow$ Besoin de variables pour interpréter ces dimensions de variabilité

Interprétation du graphe des individus grâce aux variables

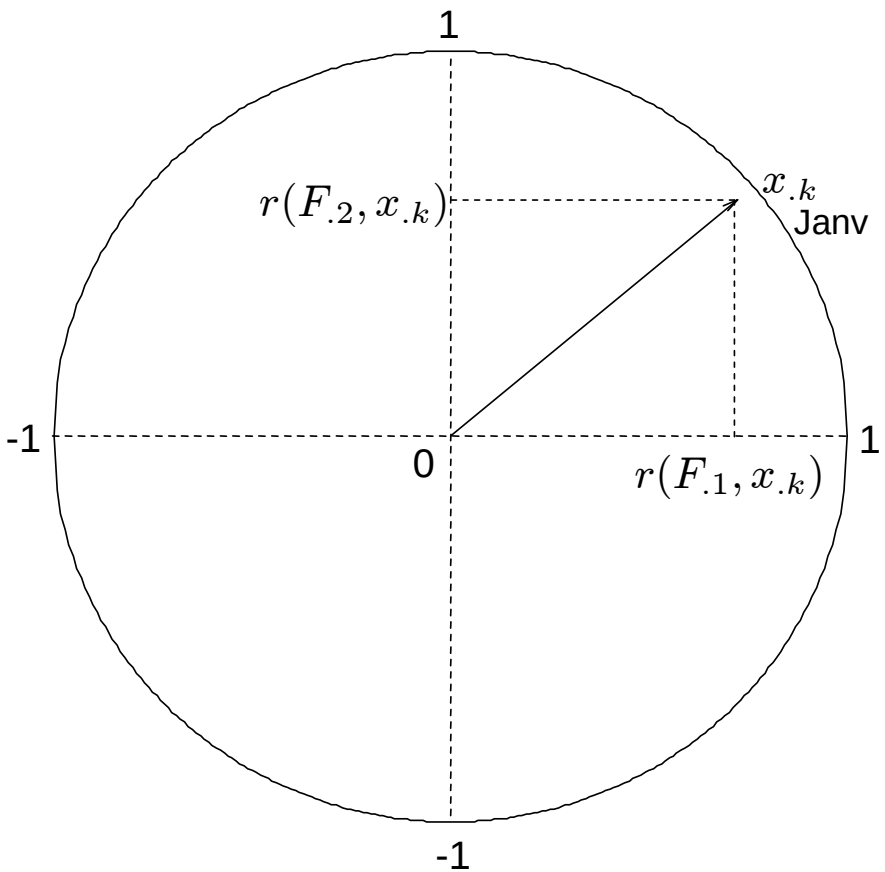
Considérons les coordonnées des individus sur les axes comme des variables





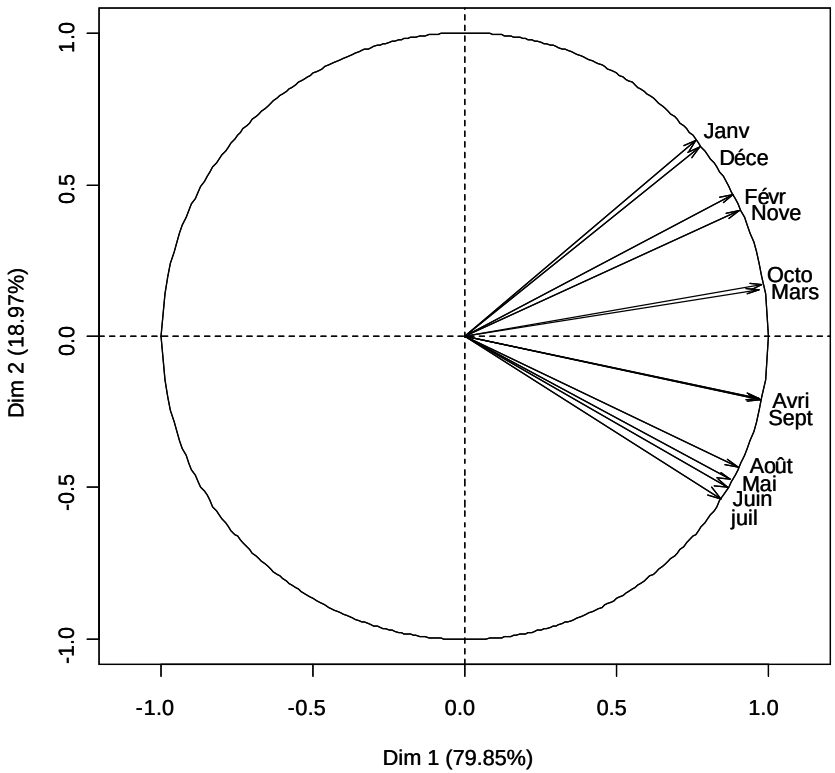
Interprétation du graphe des individus grâce aux variables

- Corrélations entre la variable $x_{.k}$ et $F_{.1}$ (et $F_{.2}$)



$\Rightarrow$ Cercle des corrélations

Interprétation du graphe des individus grâce aux variables



Toutes les variables sont corrélées à F_1.
Comment interpréter le 1er axe ?

Comment interpréter le 2ème ?

Principaux facteurs de variabilité :

1 - villes chaudes et froides ;

2 - à T^{o} moyenne constante : l'amplitude thermique

Analyse en Composantes Principales (ACP)

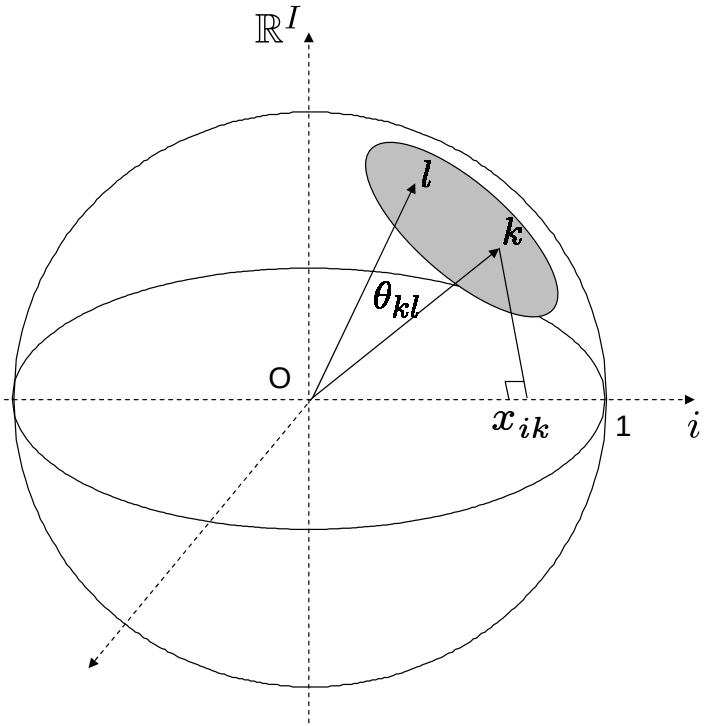
1 Données - Exemples

2 Etude des individus

3 Etude des variables

4 Aides à l'interprétation

Nuage des variables N^K



1 variable = 1 point dans un espace à I dimensions

$$\cos(\theta_{kl}) = \frac{< x_{.k}, x_{.l} >}{\|x_{.k}\| \ \|x_{.l}\|} = \frac{\sum_{i=1}^{I} x_{ik} x_{il}}{\sqrt{\sum_{i=1}^{I} x_{ik}^2} \sqrt{\sum_{i=1}^{I} x_{il}^2}}$$

Comme les variables sont centrées : $\cos(\theta_{kl}) = r(x_{.k}, x_{.l})$

Si variables réduites $\Rightarrow$ points sur une hypersphère de rayon 1

Ajustement du nuage des variables

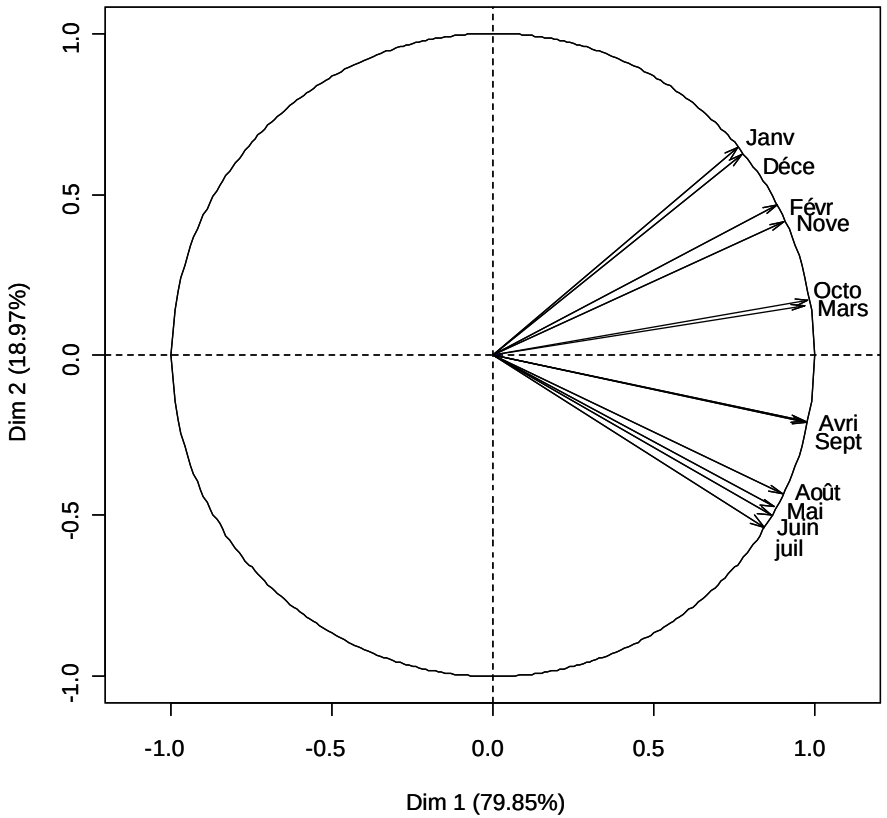
Même règle que pour les individus : recherche d'axes orthogonaux

$$\underset{v_1 \in \mathbb{R}^I}{\arg\max} \sum_{k=1}^{K} r(v_1, x_{.k})^2$$

$\Rightarrow$ v_1 est la variable synthétique qui résume au mieux les variables

Trouver le $2^{\text{ème}}$ axe, puis le $3^{\text{ème}}$, etc.

Ajustement du nuage des variables

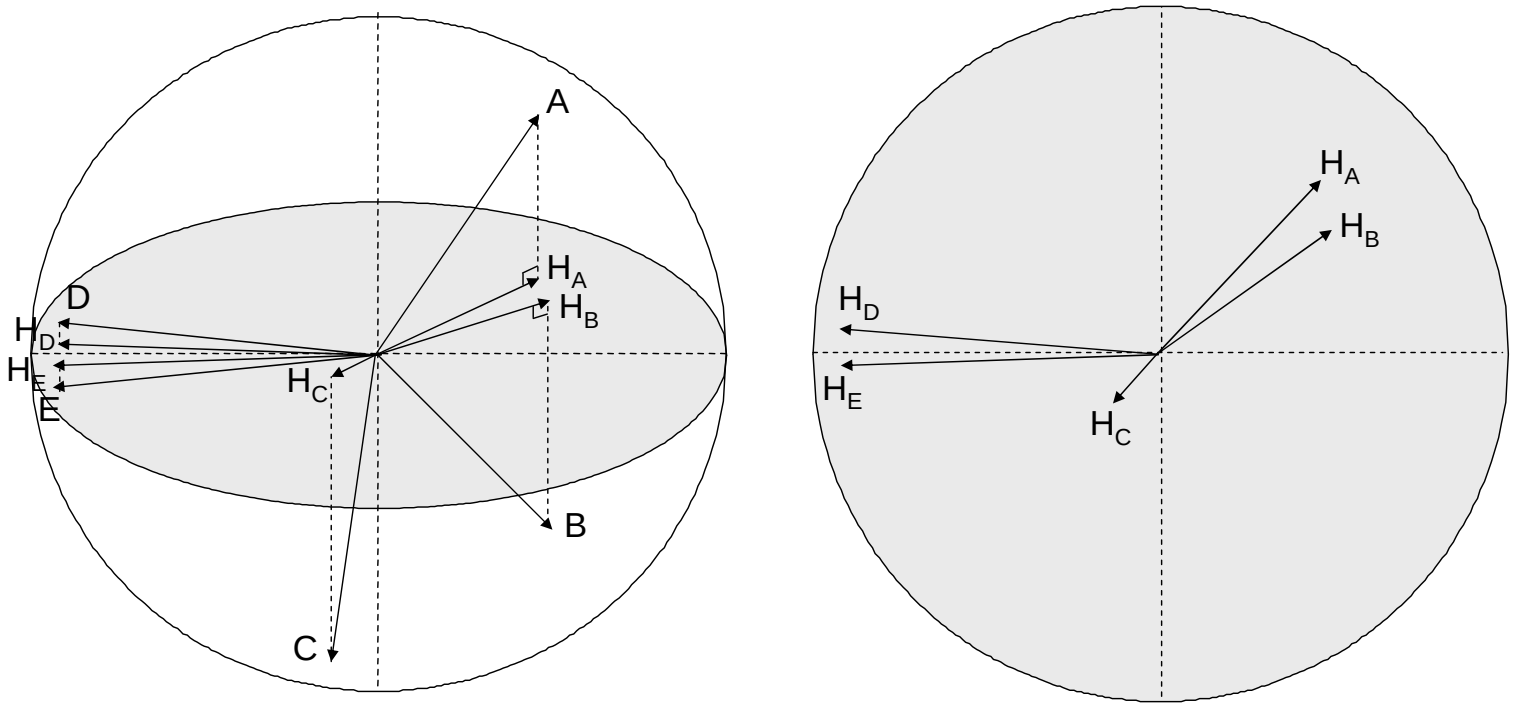


$\Rightarrow$ Même représentation que précédemment ! ! ! !

- aide pour interpréter les individus
- représentation optimale du nuage des variables
- visualisation de la matrice des corrélations

Projections...

$r(A, B) = cos(\theta_{A,B})$
$cos(\theta_{A,B}) \approx cos(\theta_{H_A,H_B})$ si les variables sont bien projetées



Seules les variables bien projetées peuvent être interprétées !

Analyse en Composantes Principales (ACP)

1 Données - Exemples

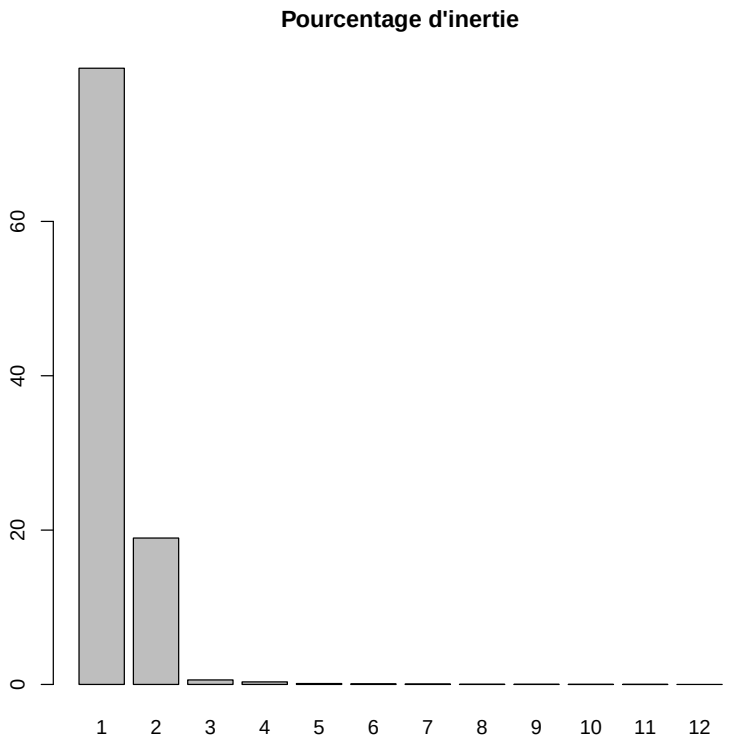
2 Etude des individus

3 Etude des variables

4 Aides à l'interprétation

Pourcentage d'inertie

- Pourcentage d'information (d'inertie) expliqué par chaque axe



$\Rightarrow$ Choix d'un nombre de dimensions à interpréter

Pourcentage d'inertie si indépendance entre variables

	Nombre de variables												
nbind	4	5	6	7	8	9	10	11	12	13	14	15	16
5	96.5	93.1	90.2	87.6	85.5	83.4	81.9	80.7	79.4	78.1	77.4	76.6	75.5
6	93.3	88.6	84.8	81.5	79.1	76.9	75.1	73.2	72.2	70.8	69.8	68.7	68.0
7	90.5	84.9	80.9	77.4	74.4	72.0	70.1	68.3	67.0	65.3	64.3	63.2	62.2
8	88.1	82.3	77.2	73.8	70.7	68.2	66.1	64.0	62.8	61.2	60.0	59.0	58.0
9	86.1	79.5	74.8	70.7	67.4	65.1	62.9	61.1	59.4	57.9	56.5	55.4	54.3
10	84.5	77.5	72.3	68.2	65.0	62.4	60.1	58.3	56.5	55.1	53.7	52.5	51.5
11	82.8	75.7	70.3	66.3	62.9	60.1	58.0	56.0	54.4	52.7	51.3	50.1	49.2
12	81.5	74.0	68.6	64.4	61.2	58.3	55.8	54.0	52.4	50.9	49.3	48.2	47.2
13	80.0	72.5	67.2	62.9	59.4	56.7	54.4	52.2	50.5	48.9	47.7	46.6	45.4
14	79.0	71.5	65.7	61.5	58.1	55.1	52.8	50.8	49.0	47.5	46.2	45.0	44.0
15	78.1	70.3	64.6	60.3	57.0	53.9	51.5	49.4	47.8	46.1	44.9	43.6	42.5
16	77.3	69.4	63.5	59.2	55.6	52.9	50.3	48.3	46.6	45.2	43.6	42.4	41.4
17	76.5	68.4	62.6	58.2	54.7	51.8	49.3	47.1	45.5	44.0	42.6	41.4	40.3
18	75.5	67.6	61.8	57.1	53.7	50.8	48.4	46.3	44.6	43.0	41.6	40.4	39.3
19	75.1	67.0	60.9	56.5	52.8	49.9	47.4	45.5	43.7	42.1	40.7	39.6	38.4
20	74.1	66.1	60.1	55.6	52.1	49.1	46.6	44.7	42.9	41.3	39.8	38.7	37.5
25	72.0	63.3	57.1	52.5	48.9	46.0	43.4	41.4	39.6	38.1	36.7	35.5	34.5
30	69.8	61.1	55.1	50.3	46.7	43.6	41.1	39.1	37.3	35.7	34.4	33.2	32.1
35	68.5	59.6	53.3	48.6	44.9	41.9	39.5	37.4	35.6	34.0	32.7	31.6	30.4
40	67.5	58.3	52.0	47.3	43.4	40.5	38.0	36.0	34.1	32.7	31.3	30.1	29.1
45	66.4	57.1	50.8	46.1	42.4	39.3	36.9	34.8	33.1	31.5	30.2	29.0	27.9
50	65.6	56.3	49.9	45.2	41.4	38.4	35.9	33.9	32.1	30.5	29.2	28.1	27.0
100	60.9	51.4	44.9	40.0	36.3	33.3	31.0	28.9	27.2	25.8	24.5	23.3	22.3

TABLE: Quantile à 95 % du pourcentage d'inertie des 2 premières dimensions de 10000 PCA obtenue avec des variables indépendantes

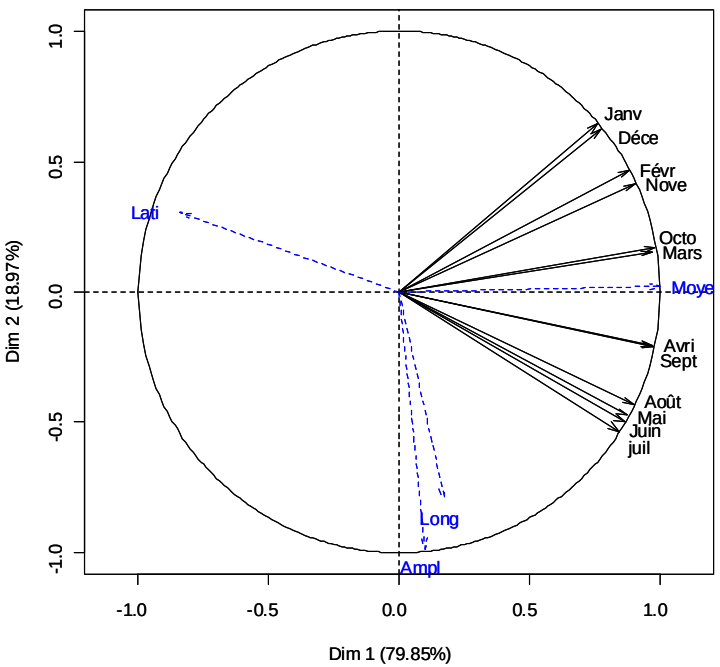
Pourcentage d'inertie si indépendance entre variables

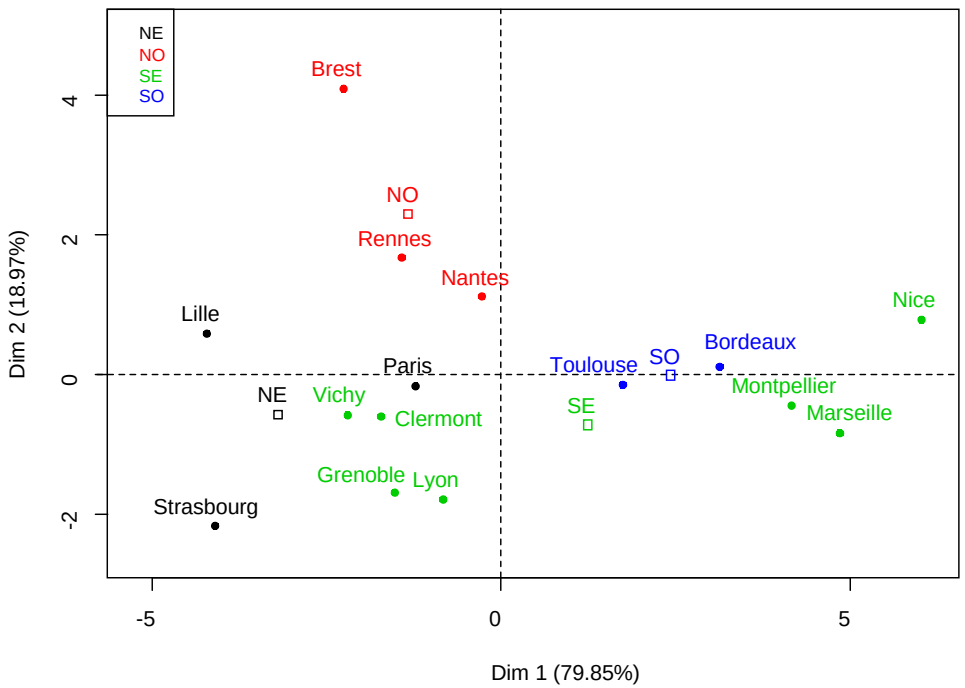
	Nombre de variables												
nbind	17	18	19	20	25	30	35	40	50	75	100	150	200
5	74.9	74.2	73.5	72.8	70.7	68.8	67.4	66.4	64.7	62.0	60.5	58.5	57.4
6	67.0	66.3	65.6	64.9	62.3	60.4	58.9	57.6	55.8	52.9	51.0	49.0	47.8
7	61.3	60.7	59.7	59.1	56.4	54.3	52.6	51.4	49.5	46.4	44.6	42.4	41.2
8	57.0	56.2	55.4	54.5	51.8	49.7	47.8	46.7	44.6	41.6	39.8	37.6	36.4
9	53.6	52.5	51.8	51.2	48.1	45.9	44.4	42.9	41.0	38.0	36.1	34.0	32.7
10	50.6	49.8	49.0	48.3	45.2	42.9	41.4	40.1	38.0	35.0	33.2	31.0	29.8
11	48.1	47.2	46.5	45.8	42.8	40.6	39.0	37.7	35.6	32.6	30.8	28.7	27.5
12	46.2	45.2	44.4	43.8	40.7	38.5	36.9	35.5	33.5	30.5	28.8	26.7	25.5
13	44.4	43.4	42.8	41.9	39.0	36.8	35.1	33.9	31.8	28.8	27.1	25.0	23.9
14	42.9	42.0	41.3	40.4	37.4	35.2	33.6	32.3	30.4	27.4	25.7	23.6	22.4
15	41.6	40.7	39.8	39.1	36.2	34.0	32.4	31.1	29.0	26.0	24.3	22.4	21.2
16	40.4	39.5	38.7	37.9	35.0	32.8	31.1	29.8	27.9	24.9	23.2	21.2	20.1
17	39.4	38.5	37.6	36.9	33.8	31.7	30.1	28.8	26.8	23.9	22.2	20.3	19.2
18	38.3	37.4	36.7	35.8	32.9	30.7	29.1	27.8	25.9	22.9	21.3	19.4	18.3
19	37.4	36.5	35.8	34.9	32.0	29.9	28.3	27.0	25.1	22.2	20.5	18.6	17.5
20	36.7	35.8	34.9	34.2	31.3	29.1	27.5	26.2	24.3	21.4	19.8	18.0	16.9
25	33.5	32.5	31.8	31.1	28.1	26.0	24.5	23.3	21.4	18.6	17.0	15.2	14.2
30	31.2	30.3	29.5	28.8	26.0	23.9	22.3	21.1	19.3	16.6	15.1	13.4	12.5
35	29.5	28.6	27.9	27.1	24.3	22.2	20.7	19.6	17.8	15.2	13.7	12.1	11.1
40	28.1	27.3	26.5	25.8	23.0	21.0	19.5	18.4	16.6	14.1	12.7	11.1	10.2
45	27.0	26.1	25.4	24.7	21.9	20.0	18.5	17.4	15.7	13.2	11.8	10.3	9.4
50	26.1	25.3	24.6	23.8	21.1	19.1	17.7	16.6	14.9	12.5	11.1	9.6	8.7
100	21.5	20.7	19.9	19.3	16.7	14.9	13.6	12.5	11.0	8.9	7.7	6.4	5.7

TABLE: Quantile à 95 % du pourcentage d'inertie des 2 premières dimensions de 10000 PCA obtenue avec des variables indépendantes

Information supplémentaire

- Pour les variables quantitatives : projection des variables
- Pour les modalités : projection au barycentre des individus qui prennent cette modalité





$\Rightarrow$ Information supp. ne participe pas à la construction des axes

Qualité de représentation – contribution

- Qualité de représentation d'une variable et d'un individu

$\cos^2$ entre une var. et sa projection

```
round(res.pca$var$cos2,2)
      Dim.1 Dim.2 Dim.3
Janv   0.58  0.42  0.00
Févr   0.78  0.22  0.00
```

$\cos^2$ entre Oi et OH_i

```
round(res.pca$ind$cos2,2)
          Dim.1 Dim.2 Dim.3
Bordeaux   0.95  0.00  0.05
Brest      0.23  0.76  0.00
```

$\Rightarrow$ Seuls les éléments bien projetés peuvent être interprétés

- Contribution d'1 var. et d'1 individu à la construction de l'axe s :

$$Ctr_s(k) = \frac{r(x_{.k}, v_s)^2}{\sum_{k=1}^{K} r(x_{.k}, v_s)^2} \ (\times 100) \qquad Ctr_s(i) = \frac{F_{is}^2}{\sum_{i=1}^{I} F_{is}^2} \ (\times 100)$$

```
round(res.pca$var$contrib,2)
     Dim.1 Dim.2 Dim.3
Janv  6.05 18.24  0.66
Févr  8.09  9.67  1.61
```

```
round(res.pca$ind$contrib,2)
          Dim.1 Dim.2 Dim.3
Bordeaux   6.78  0.03 49.48
Brest      3.58 49.07  1.26
```

$\Rightarrow$ Eléments avec une forte coordonnée contribuent le plus

Description des dimensions

Par les variables quantitatives :

- calcul des corrélations entre chaque variable et la dimension *s*
- tri des coefficients de corrélation (significatifs)

```
> dimdesc(res.pca)
$Dim.1$quanti
     correlation      p.value
Moye   0.9997097 0.000000e+00
Octo   0.9801599 1.609672e-10
Sept   0.9740289 9.130414e-10
Avri   0.9693357 2.657670e-09
Mars   0.9687704 2.988670e-09
Nove   0.9037531 3.834950e-06
...
juil   0.8415346 8.385040e-05
Déce   0.7743349 7.017832e-04
Janv   0.7612384 9.784512e-04

Lati  -0.8389348 9.259113e-05
```

```
$Dim.2$quanti
     correlation      p.value
Janv   0.6443379 9.519348e-03
Déce   0.6242957 1.285835e-02

juil  -0.5314197 4.148657e-02
Long  -0.7922192 4.298867e-04
Ampl  -0.9856753 1.963381e-11
```

Description des dimensions

Par les variables qualitatives :

- Analyse de variance des coordonnées des individus sur l'axe s (variable Y) expliqués par la variable qualitative
  - un test F par variable
  - un test t de Student par modalité pour comparer la moyenne de la modalité avec la moyenne générale

```
> dimdesc(res.pca)
$Dim.2$quali
              R2    p.value
Région 0.6009012 0.01467946

$Dim.2$category
     Estimate     p.value
NO  2.0503647 0.003530801
SE -0.9738852 0.047120253
```

Pratique de l'ACP

1. Choisir les variables actives
2. Choisir de réduire ou non les variables
3. Réaliser l'ACP
4. Choisir le nombre de dimensions à interpréter
5. Interpréter simultanément le graphe des individus et celui des variables
6. Utiliser les indicateurs pour enrichir l'interprétation
7. Revenir aux données brutes pour interpréter

Suppléments

Analyse de données avec R (2009).
Husson, Lê, Pagès.
Presses Universitaires de Rennes (15 Euros)

Package `FactoMineR` pour faire des ACP :
`http://factominer.free.fr/index_fr.html`

Vidéos sur Youtube :

- une chaîne Youtube : `youtube.com/HussonFrancois`
- une playlist de vidéos en français
- une playlist de vidéos en anglais